**CGIAR Research Program on**
**Climate Change, Agriculture and Food Security (CCAFS)**

# CCAFS Dataverse
# *Video Transcript*

**October 2013**



These materials were produced for and with funding from the Climate Change Agriculture and Food Security Research Program of the Consultative Group on International Agricultural Research (CGIAR).

## Introduction

The CCAFS Dataverse has been set up to house the data and documentation and other files from the baseline surveys. This demonstration shows the structure of the CCAFS database which currently is hosted at Harvard University. There are other videos available that provide a general introduction to Dataverse and show how to register as a user and create your own Dataverse.

## Structure of the CCAFS Dataverse

The CCAFS Dataverse has been set up with a main collection – CCAFS Baseline Surveys – and 3 sub-collections – one each for the Household Baseline, Village Baseline and Organisational Baseline.

We will concentrate on the Household Baseline in this demonstration.

Within Dataverse you can group the study files into categories – the categories and the files within the categories will generally be ordered alphabetically. In the CCAFS Dataverse we have numbered the categories and some of the files so that they appear sorted in a more logical way.

Thus we have

01. The core questionnaires – including translated copies of the questionnaire
02. Added Questionnaire modules – for example the livestock module used in Ethiopia
03. Manuals – fieldworker training manual, data checking manual, etc.
04. Training videos – how to complete parts of the questionnaire and data entry system – some local teams had problems understanding some of the concepts we were using the survey – such as what we mean by "Not applicable" so we created a few training videos for use in new sites.
05. Data entry system – this was created in CS-Pro and under this category we have a zip file containing the CS-Pro files for the latest version of the system. Also in this category we have included versions of the system that have been adapted for particular sites. For example we have a version where the data entry screens are in Spanish for use in Nicaragua and Costa Rica and another version used in Mozambique where some of the screen text has been translated into Portuguese.
06. Process reports – this section includes a document described the recoding and merging process for consolidating livestock and crop codes
07. Analysis reports – these are the site reports
08. Data files – loaded as SPSS files and converted by Dataverse into tabular format enabling users to run some summaries online. There are two files for the 15 core sites – the household level data and the fertiliser level data; and we also have two files for any additional sites.
09. Syntax files – these are SPSS syntax files for labelling and for the initial analysis
10. CSPro data and analysis outputs for core sites in East Africa – the CSPro data files have restricted access because they contain identification information
11. Original data and analysis outputs for West Africa sites
12. Similar files for the core sites in South Asia

13. Original data and analysis outputs for additional sites – files for any new sites will be added here as and when they become available. Currently this section includes data for the extra sites from India, Bangladesh & Nepal and more recent sites from Mozambique, Costa Rica and Nicaragua.
14. Sampling Frames – also restricted – you can restrict the whole study or individual files. There is an option that allows users to request access – we are no longer using that option but are just allowing access to specified users from the CCAFS team.

## Read Me files

The first category includes a ReadMe file and a Data Quality Summary which describes the structure of the Dataverse and points out any issues with the data. We include a note in the study description recommending users read these files first before downloading and using the data. The first time anyone downloads a file in a particular session they are asked to complete a guestbook and agree to the terms and conditions.

## Guest Book

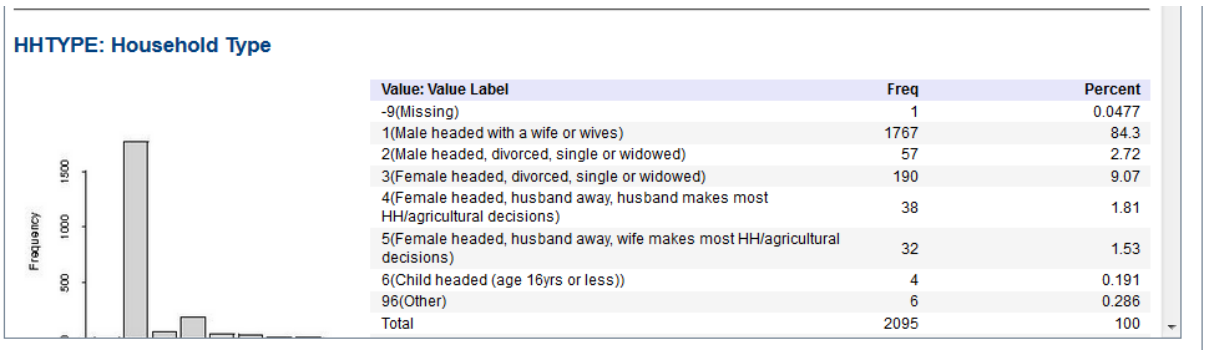
Dataverse includes an optional guest book feature. This enables the owners of the Dataverse to keep track of who is downloading files from the studies. The guest book form appears the first time a user downloads a file or uses the analysis option on a data file. The guest book automatically takes the user's name, email, institution and position from his/her login details and you can add extra fields to gather additional information. We just have an open text field for users to indicate how they are intending to use the data.

## Data Files

The data files were uploaded as SPSS files and during the upload Dataverse converted them into Tabular Data. This process means the file can be downloaded in a variety of formats – SPSS, Stata, R, etc.

You can also do some brief analysis online. This is generally just frequency tables and summary statistics but can be useful to get an overall feel for the data.

For example we can very quickly look at the distribution of household type and household size. For these discrete variables we can display a frequency table and bar chart.



**HHTYPE: Household Type**




| Value: Value Label | Freq | Percent |
|---|---|---|
| -9(Missing) | 1 | 0.0477 |
| 1(Male headed with a wife or wives) | 1767 | 84.3 |
| 2(Male headed, divorced, single or widowed) | 57 | 2.72 |
| 3(Female headed, divorced, single or widowed) | 190 | 9.07 |
| 4(Female headed, husband away, husband makes most HH/agricultural decisions) | 38 | 1.81 |
| 5(Female headed, husband away, wife makes most HH/agricultural decisions) | 32 | 1.53 |
| 6(Child headed (age 16yrs or less)) | 4 | 0.191 |
| 96(Other) | 6 | 0.286 |
| Total | 2095 | 100 |

For continuous variables such as land values the system would produce a histogram, normal probability plot and summary statistics.

## Cataloguing Information

Each study you create should have a set of cataloguing information – this is part of the meta-data for the study. In Dataverse the only compulsory elements in the catalogue are the title and the study ID. The study ID is a unique identifier which must be set when the study is created and cannot be changed as it then forms part of the unique citation.

The title, along with all other elements in the catalogue can be edited at later stages though we would recommend trying to complete as much of this information as possible when you create the study. We have a separate written guide detailing each of the cataloguing elements. Also we recommend you look at our meta-data guide where we recommend gathering the study catalogue information together at the start of the project, updating it as appropriate during the project. This makes the process of archiving much simpler as you have more complete information when you come to create the Dataverse.

## Archive Backup

As a sort of backup to the Dataverse we have a local storage device where we keep copies of the files currently on the Dataverse. The folder structure on this local device mimics the structure of the Dataverse. So for example under the Household Baseline Survey we have folders that match the file categories used in the study. This procedure of having both a local archive and a public archive is not strictly necessary but can be useful particularly if you have some files that you are not putting in the public domain but still want/need to keep copies of. A local copy can also be reassuring – we might refer to this as a "belt and braces" approach.

## Summary

Dataverse is just one example of an archiving system that we have found relatively easy to use. Whatever system you choose for your archive we would advice you do the following:

- Start to gather your study cataloguing information together from the start of your project;
- Think about the structure of your archive;
- Keep a local archive where the structure matches the structure of the online archive.

# Appendix I – CCAFS Data Management Support Pack

This document is part of the CCAFS Data Management Support Pack produced by the Statistical Services Centre, University of Reading, UK. The following materials are available in the pack:

0. Data Management Strategy
    a. CCAFS Data Management Strategy
1. Research Protocols
    a. Writing Research Protocols – a statistical perspective
    b. Preparation of Research Protocols – Good Practice Case Study
    c. What is a Research Protocol, and how to use one (Video & Transcript)
    d. Details of what a Research Protocol should contain (Video & Transcript)
2. Data Management Policies & Plans
    a. Creating a Data Management Plan
    b. Data Management Plan (Video & Transcript)
    c. Example Data Management Activity Plan
    d. Example Consent Form
3. Budgeting & Planning
    a. Budgeting & Planning for Data Management
    b. ToR Data Support Staff
    c. Budgeting & Planning (Video & Transcript)
4. Data Ownership
    a. Data Ownership and Authorship
    b. Template – Data Ownership Agreement
    c. CCAFS Data Ownership & Sharing Agreement
    d. Data Ownership & Authorship (Video & Transcript)
5. Data & Document Storage
    a. Creating and Using a DDS
    b. DDS Introduction – (Video & Transcript)
    c. DDS Organisation – (Video & Transcript)
    d. DDS Ownership – (Video & Transcript)
    e. Introduction to Dropbox – (Video & Transcript)
6. Archiving & Sharing
    a. Archiving & Sharing Data
    b. Data and Documents to Submit for Archiving – a checklist
    c. MetaData
    d. Archiving & Sharing (Video & Transcript)
    e. Metadata (Video & Transcript)
    f. CCAFS HBS Questionnaire
    g. CCAFS HHS Code Book
    h. CCAFS Training Manual for Field Supervisors

7. CCAFS Data Portals
    a. Portals for CCAFS Outputs
    b. AgTrials Summary
    c. CCAFS-Climate Summary
    d. DSpace Introduction
    e. Introduction to Dataverse (Video & Transcript)
    f. Creating a Dataverse (Video & Transcript)
    g. Dataverse Study Catalogue
    h. CCAFS Dataverse (Video & Transcript)
8. Data Quality & Organisation
    a. Data Quality Assurance
    b. Guidance for handling different types of Data
    c. Transition from Raw to Primary Data
    d. Data Quality Assurance (Video & Transcript)
    e. Guidance for handling different types of data (Video & Transcript)
    f. Transition from Raw to Primary Data (Video & Transcript)